**Intro**

The purpose of this project was to gain exposure to the extraction transformation and loading of data into a SQL database, as well as the creation of the database and the ability to query through the database to answer specific questions. The domain of this data was set around understanding the effects of temperature and humidity on the spread of the COVID-19 virus.

**Data Mining**

**raw\_data**

This required 2 sets of data to be obtained. The weather data was downloaded from weathersource.com, which gave day by day date, temperature (min, avg, max,), humidity (min, avg, max,), and fips code (county id number). The COVID-19 data was found off of the “CDC COVID Tracker” website, and gave date, fips code, new cases and deaths, along with several other variant ratios to population. Google Cloud Platform’s “BigQuery engine” was used to obtain api request from weather source.com. Because there was a data limit, I only requested the month of July 2020 (96,000 entries). I read in and parsed both sets of data in python 3.

**Database Creation**

**create\_database.py**

The database was created in SQLite 3.

**Data Cleaning**

**filter\_covid\_cases\_by\_date.py**

**populate\_database.py**

The date format along with several null values were cleaned before the set was ready to be loaded into SQL. The database was created in SQLite 3.

**Database Queries**

**query\_examples**

Query 1 was a simple join on both tables to make sure that the 1 to 1 ratio of the fips code showed no mismatching values or duplicates. Output file = test.csv

Query 2 was to find fips codes which contained large amounts of new cases. Output file = test1.csv

Query 3 was to find the counties with the coldest weather by fips code. Output file = test2.csv

Query 4 was to find fips codes that have higher number of covid cases with lower temperature, using a set range. I tried to find the standard distribution to obtain my Q1 and Q3 outliers (covid\_distrib.png). The Q1 was around 50. So that was set as the minimum temperature.

Query 5 was to find the fips codes with high new covid cases which also exceeded average humidity range. Found to be around 15.

**Conceptual model**

**preliminary\_conceptual\_model.png**

The preliminary conceptual model was created using lucidapp from Lucid Charts.

**Discussion**

This project turned out to be a success. It help me gain a greater understanding of cleaning data in python as well as creating a database and making queries within the database. It also showed me the value of these skills with real world problems and applicable research. If I were to add to the research of this domain, I would set up a machine learning script to draw significance between these 2 data sets, since there were too many variables to interpret no matter how specific I could make the SQL queries. If this project were to be conducted again I would pick a dataset such as political voting data to conjoin with my covid data. This would be much easier to query and draw a conclusion.